Functional Document Testing

- Michael Stahl, Red Hat, Inc.
- 2016-09-08

Functional Document Testing

1) Problem Statement / Requirements

2) Diff Basics

3) Non-Determinism

4) First Results

Problem: Change Is Bad

Data-loss regressions in ODF round trips like

- Tdf#74230 graphic defaults changed
- Tdf#92379 background properties lost
- Tdf#100182 document index marks lost

Status Quo

Automated Crash Testing:

- 90k documents automatically loaded & stored
- success = LO doesn't crash
- no check of the content of written documents
- Idea: compare written documents to "reference" docs

  (at least for ODF round-trips, the most important format)

Non-Requirements

General purpose ODF diff tool

- Different producers
- User initiated content changes

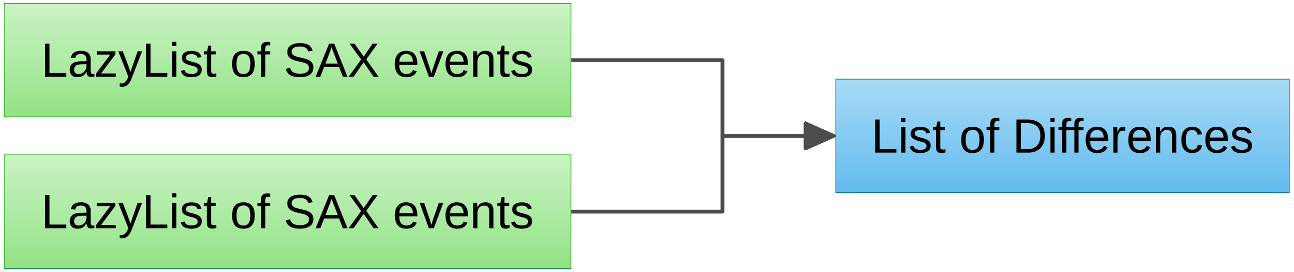
Requirements

- Scalable: we have 27k documents (threading...)
- Performance: want to compare faster than LO generates
- Memory usage: DOMs wasteful?
- Can make assumptions about documents: written by LO
  - UTF-8
  - Namespaces
- Non-determinism
- Intentional changes

Putting the Fun in Functional

How to compare 2 XML documents?

- save memory => avoid DOM



but: SAX callback interface? Haskell

- could Greenspun some CPS transform ...
- ... or be pragmatic and use a language where this is idiomatic

Highly Unsophisticated Algorithm

diff :: State → [Event] → [Event] → [Difference]

- compare first SAX events in both lists
- if they match, continue with rests of both lists
- if not:
  - cut 5 elements from both lists
  - try to find shortest "edit distance" between them (smallest number of insert / delete / difference)
  - report differences
  - continue with the first elements that match again

Non-Determinism: ODF

ODF has non-deterministic features:

- Spreadsheet formulas like `RAND`, `NOW`, `INFO`
- Presentations: random animations
- Fields: `<text:date>` / `<text:time>` etc.
- `meta.xml` `<dc:date>` / `<meta:creation-date>` (templates)
- XForms bindings: `now()`

Non-Determinism: Application Specific

- Automatic Styles
- Writer layout: how far did it go?
- Various generated “`*:id`” attributes
- ToX-mark / Reference-mark order
- ...

To Fix LO or to Ignore?

Reasons for fixing LO to be deterministic:

- Non-determinism is annoying for a small sub-set of users

Reasons for ignoring non-determinism in ODFunDiff:

- Inherent ODF non-determinism
- Intentional changes
- Older releases
- Want to compare production releases, not special “mode”

Automatic Styles

- Cannot compare directly (non-determinism)
- Cut them out of the document
- Usage (attributes of type `styleNameRef`) creates a mapping
- At end of document:
  - use mapping to compare style content
  - compare unused automatic styles

Stateless Pre-Filtering

```
preFilter :: [Event] → [Event]
```

- Eliminate `<text:soft-page-break>` element (layout)
  - Merge surrounding `<text:span>` / `<text:a>` / `<text:s>`
- Sort ToX-marks / Ref-marks
  - such that the IDs can be mapped later...
- Sort `<config:config-item>`
- Sort `<manifest:file-entry>`
  - … and remove "Configurations2/accelerator/current.xml"
  - … and remove “layout-cache”
- Merge multiple consecutive character data events

Stateful Attribute Value Mapping

- Generated ID attributes
  - `styleNameRef` attributes with automatic styles
  - `<text:list>` `xml:id` / `text:continue-list`
  - ToX mark start / end `text:id`
  - `<text:changed-region>` `text:id` / `<text:change>` `text:change-id`
  - Random animations: `draw:id`, `smil:targetElement`, `smil:begin`

Missing Attributes

- Writer table `style:width` / `style:column-width` (layout)
- Writer `text:use-soft-page-breaks` (layout)
  - Most pointless attribute ever

Attribute Value and Character Data Changes

- Writer `draw:z-index` (layout)
- `officeooo:rsid` (generated)
- Fields: `<text:page-number>`, `<text:page-count>` (layout)
- Header: `<text:chapter>`, `<text:variable-get>` (layout)
- Footnotes: `<text:note-citation>` (layout)
- `<meta:document-statistic>` (layout, async. word-count)
- `<meta:generator>`
- Config-item “DoNotCaptureDrawObjsOnPage” (layout)

Teach Your Tool Ignorance With a Big Hammer

- Non-localized changes:
  - Spreadsheet formulas like `RAND`, `NOW`
  - Random animations
- Regex search `content.xml` => pass in flag to ignore more stuff

Are We Sufficiently Ignorant Yet?

- If written by same LO build, around 20 documents still differ
- ooo71392-1.ods spreadsheet with randomized styles
- tdf89783-12.odt 64k automatic table styles => OOM

It’s Not a Bug, it’s a Feature

- Filter differences that are intentional changes
- Implemented as separate result filter function

```
postFilter :: [Difference] → [Difference]
```

- Version dependent
  - Ignore change if reference version < N and test version >= N
  - For 5.3, ~15 changes identified

Bugs Found

- Hyperlinks on `<form:form>` not converted to absolute URLs
- OOo 1.0 XML: Impress converting legacy animations via timer
- Calc `ISOWEEKNUM` migration issue
- Number format “`month`” vs. “`minutes`” regression
- Calc wrong matrix formula result “Err:504”
- Calc header / footer lost?
- some other number formatting bug
- Calc `CEILING` formula not returning an error (now declared a feature)

Surely Functional Means Slow?

- GHC profiling helps
- LO release build round-trips documents in 90 minutes
- ODFunDiff compares them in 18 minutes
  - Trivial multi-threading with GHC runtime and Chan

Thank you …

- git://gerrit.libreoffice.org/odfundiff